Embedding Learning Through Multilingual Concept Induction

Philipp Duffer, Mengjie Zhao, Martin Schmitz, Alexander Fraser, Hinrich Schütze
1 Center for Information and Language Processing (CIS) LMU Munich, Germany
2 École Polytechnique Fédérale de Lausanne, Switzerland
{philipp, martin, fraser}@cis.lmu.de, mengjie.zhao@epfl.ch

Motivation

- Multilingual wordspace with 1000s (low-resource) languages.
- Objective: Multilingual wordspace with 1000s (low-resource) languages.
- New feature: concept-IDs.

Dictionary Induction

- Creating ChI dictionaries for n-gram languages.
- Computing \( p(n) \) by aggregating fast-align alignments.

Concept Induction

- CLIQUE Projection
- In ideal world: concepts correspond to ciphers in the multilingual dictionary graph.
- In real-world: identify quasi-ciphers to accommodate noise.
- Identity concepts in pivot dictionary graph and project them onto target languages.

Dictionary Baselines

- Non-resource-based: RTSIMPLE: Roundtrip translation directly on the dictionary graph.
- Context-based baselines: bag-of-words (BOW) [Vaclav et al., 2015], sentence-ID (S-ID) [Levy et al., 2017].
- Alternative concept-identification method: SAMPLE [Lardière et al., 2009]. Sampling-based approach for identifying concepts in a sentence-aligned corpus. Projection step same as for CLIQUE.

Baseline Results

- creation of English silver standard using the Vader-Classifier in combination with manual annotations.
- Creation of English silver standard using the Vader-Classifier in combination with manual annotations.
- Creation of English silver standard using the Vader-Classifier in combination with manual annotations.

Evaluation

Roundtrip translation

- 70 English queries taken from a list of universal words by Swedish (1946).
- Bind and released groundtruth: \( G(q) \) contains words with the same lemma as \( q \).
- Accuracy computed by \( \frac{1}{|G(q)|} \sum_{G} \frac{|G|}{|G^{*}|} \), aggregated over queries.
- \( |G^{*}| \) is varied as follows: \( S^1 = 1.1 \), \( S^4 = 2 \), \( S^6 = 2 \), and \( R^1 = 1.1 \).

Sentiment Analysis

- "How is life salvation... the power of his Christ: for the accuse... cast down, which accused them before our God..."
- Creation of English silver standard using the Vader-Classifier in combination with manual annotations.
- Creation of English silver standard using the Vader-Classifier in combination with manual annotations.
- Creation of English silver standard using the Vader-Classifier in combination with manual annotations.

Results

- roundtrip translation: 2.18
- sentence analysis: 0.84

We gratefully acknowledge funding from the European Research Council (grants 740516 & 640550) and through a Zentrum Digitalisierung.Bayern fellowship awarded to the first author. We are indebted to Michael Cyousse for making PSS available to us.

Conclusions

1. Concept-based methods outperform previous approaches.
2. New roundtrip evaluation is an excellent wordspace quality indicator.
3. Character-level is better than word-level for sentiment classification.